DEIM Forum 2019 E1-4

Dialogue Quality Distribution Prediction based on a Loss that Compares Adjacent Probability Bins

Sosuke KATO† and Tetsuya SAKAI†

† Department of Computer Science and Engineering, Waseda University
3–4–1 Okubo, Shinjuku-ku, Tokyo 169–8555, Japan
E-mail: †sow@suou.waseda.jp, †tetsuyasakai@acm.org

Abstract The NTCIR-14 Short Text Conversation Dialogue Quality subtask requires participating systems to predict a distribution of dialogue quality scores for each given customer-helpdesk dialogue. The gold distribution represents the views of multiple annotators, and the systems are evaluated by comparing the two distributions over ordinal bins of scores. In this study, we propose a loss function that is based on comparing the probabilities in adjacent bins and demonstrate its effectiveness for the task. Our proposed model outperformed the baseline model in terms of every measure for Chinese dataset.

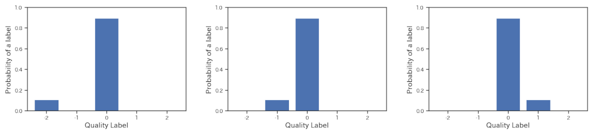
Keywords loss function, probability distribution prediction, short text conversation, dialogue quality

1 Introduction

Aiming at building a dialogue system with artificial intelligence, research on dialogue systems have received much attention, and some competitions [1,3,5] related to dialogues have been held. Tasks related to dialogues can be classified according to the property of a dialogue, e.g., a task-oriented or non-task-oriented dialogue; a human-human or human-machine dialogue, and what systems of participants do, e.g., generation, retrieval, classification, and so on.

In the End-to-End Goal-Oriented Dialog Learning Track [5] as a retrieval task, given a dialogue, a participating system selects a machine utterance or some action of the machine which should immediately follow the dialogue from the given candidates. In this task which deals with task-oriented human-machine dialogues, Precision was used as an evaluation measure and actual utterances or actions can be used as gold data.

In the case of tasks in which a system predicts some types of labels, gold labels usually are made by multiple annotators manually. However, allowing one label per item means that the gold data cannot directly represent the views of multiple annotators. In contrast, in the NTCIR-14 Short Text Conversation 3 (STC-3) [3] and the Dialogue Breakdown Detection Challenge [1], a system must predict the gold distribution which represents the views of multiple annotators. To evaluate the systems, the two distributions, i.e., the gold distribution and the predicted distribution, must be compared. The number of types of labels is relatively low, i.e., 5 in the case of STC-3, and an example of a gold distribution is shown in Figure 1 (a). When evaluating the predicted distributions like (b), (c) in Figure 1, the distance between (a) and (b) and between (a) and (c) are the same if a traditional measure such as the Jensen-Shannon divergence (JSD) is used. To solve this problem, the systems are evaluated by comparing the two distributions over ordinal bins of scores using the Normalised Match Distance (NMD) and the Root Symmetric Normalised Order-aware Divergence (RSNOD) [8] in STC-3.



(a) (b) (c)

Figure 1 Examples of distributions

In this study, we propose a loss function that is based on comparing the probabilities in adjacent bins in order to handle ordinal bins, and demonstrate its effectiveness for the NTCIR-14 STC-3 Dialogue Quality subtask.

2 Related Work

2.1 NTCIR-14 Short Text Conversation Task

The NTCIR-14 Short Text Conversation Task [3] has three subtasks; Chinese Emotional Conversation Generation subtask, Nugget Detection subtask, and Dialogue Quality (STC-3 DQ) subtask. The present study concerns the STC-3 DQ subtask, which is described below.

The STC-3 DQ subtask requires participating systems to predict a distribution of dialogue quality scores for each given customer-helpdesk dialogue. The training dataset [11] (*1) as the customer-helpdesk dialogues with annotations by multiple annotators are available and used in STC-3. The target languages of the STC-3 DQ are Chinese and English. The original language of the dialogues is Chinese and a part of dialogues was translated into English. The annotations for the STC-3 DQ subtask in this dataset are labeled in the following types of quality scores (*2);

- A-score: Task Accomplishment (Has the problem been solved? To what extent?),
- S-score: Customer Satisfaction of the dialogue (not of the product/service or the company),
- E-score: Dialogue Effectiveness (Do the utterers interact effectively to solve the problem efficiently?).

These scores are on a five-point scale: $[-2, -1, 0, 1, 2]$.

Given a dialogue and a gold distribution p^*, a system predicts a distribution p for every type of quality score and the NMD and the RSNOD as evaluation measrues are calculated.

One of the baseline model (*3) of the STC-3 DQ subtask is based on a Bidirectional Long Short-term Memory (BLSTM) [2,10] which needs a loss function to train a model. The loss function of the baseline BLSTM model are defined as follows,

$$L_{\mathrm{ce}} = -\sum_{i=1}^{B} p^*(i) \log p(i) \tag{1}$$

where B denotes the number of bins, i.e., 5 in this study.

2.2 Loss Function

In Eq. (1), Cross Entropy (CE) [6] is calculated. The CE is based on Kullback-Leibler divergence which forms the basis of JSD and therefore does not consider ordinal bins. While loss functions for classification tasks have been discussed in previous work [4], those for comparing two distributions over ordinal bins have not.

In Section 4, we proposed a loss function which considers ordinal bins and train a model using the same neural network as a baseline BLSTM model that does not use our loss function.

3 Preliminary Experiment

As a preliminary experiment, we compare the evaluation measures for ordinal bins, i.e., RSNOD and NMD, with JSD which does not consider ordinal bins. In order to compare these measures visually, we plot the heat map and show examples of distributions. First, we train the baseline BLSTM model and then we classify the gold and predicted distributions and plot the heat maps. Finally, we compare these heat maps and show some pairs of gold and predicted distributions.

3.1 Training of the Baseline Model

We divided the training dataset into the one to train the model and the one to plot the heat maps in the ratio of 8 : 2. The latter was also used as validation data. Using the divided data for training and validation, we train the baseline BLSTM model and predict distributions of the validation data to plot the heat maps.

3.2 Distribution Mapping for Classification

Focusing on considering ordinal bins, we map a distribution to a class considering adjacent probability bins using the mapping table shown in Table 1. In this mapping where adjacent probability bins are compared, distributions are mapped to 16 ($= 2^{5-1}$) types on a five-point scale. For example, the distribution of the top example in Table 1 is classified as '15' because $p(-2) \leqq p(-1)$, $p(-1) \leqq p(0)$, $p(0) \leqq p(1)$ and $p(1) \leqq p(2)$ hold in this distribution.

3.3 Differences among NMD, RSNOD, and JSD

Figures 2 to 5 visualize the A-score distributions of NMD, RSNOD, and JSD for comparison, where the x-axis represents the gold distribution classes and the y-axis represents the classes predicted by the baseline BLSTM model. The number of each cell in Figure 2 denotes the number of dialogues. The number of each cell denotes the mean NMD, RSNOD and JSD for A-score in Figures 3 to 5 respectively.

We can see from Figure 2 that most of the gold distributions are classified as '12' and '13', and the baseline BLSTM model tends to predict the distributions which are classified as '12'. Moreover, we can see over Figures 3 to 5 that distances are relatively lager when the class of the distribution is predicted as '6'.

We focus on the cells where the class of the predicted distribution is '6' and we show examples of the pairs of the gold and predicted distribution in Figures 6 and 7 where the classes of predicted distributions are '6'. In contrast, the class of gold distributions in Figure 6 is '12' and the one in Figure 7 is '6'. We also show the evaluation measures for each pair in Figures 6 and 7. Comparing Figures 6 and 7, in terms of JSD, Example 1 is considered better than Example 2, which is quite counterintuitive. In contrast, according to NMD and RSNOD, Example 2 is rated higher than Example 1. These examples suggest that we should use a loss function considering ordinal bins when NMD and RSNOD are used as the evaluation measures.

4 Proposed Method

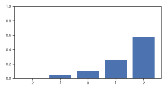
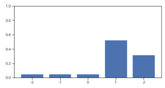
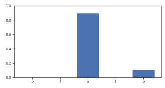
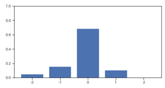
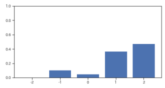
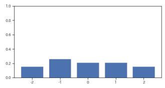
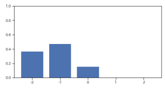
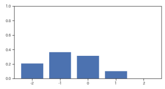
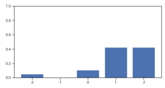
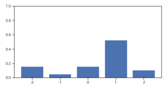
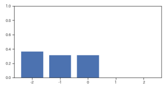
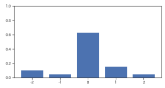
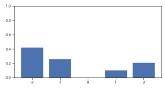
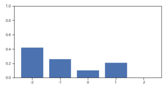
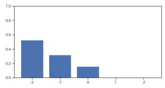
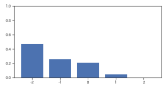
Again, from the preliminary experiment, we should use

(*1): https://sakai-lab.github.io/stc3-dataset/

(*2): http://sakailab.com/wp-content/uploads/2018/06/STC3atNTCIR-14.pdf

(*3): https://github.com/sakai-lab/stc3-baseline

Table 1 Mapping table to map a distribution p to a class

$p(-2) \leftrightarrow p(-1)$	$p(-1) \leftrightarrow p(0)$	$p(0) \leftrightarrow p(1)$	$p(1) \leftrightarrow p(2)$	class	example
$p(-2) \leqq p(-1)$	$p(-1) \leqq p(0)$	$p(0) \leqq p(1)$	$p(1) \leqq p(2)$	15	
			$p(1) > p(2)$	14	
		$p(0) > p(1)$	$p(1) \leqq p(2)$	13	
			$p(1) > p(2)$	12	
	$p(-1) > p(0)$	$p(0) \leqq p(1)$	$p(1) \leqq p(2)$	11	
			$p(1) > p(2)$	10	
		$p(0) > p(1)$	$p(1) \leqq p(2)$	9	
			$p(1) > p(2)$	8	
$p(-2) > p(-1)$	$p(-1) \leqq p(0)$	$p(0) \leqq p(1)$	$p(1) \leqq p(2)$	7	
			$p(1) > p(2)$	6	
		$p(0) > p(1)$	$p(1) \leqq p(2)$	5	
			$p(1) > p(2)$	4	
	$p(-1) > p(0)$	$p(0) \leqq p(1)$	$p(1) \leqq p(2)$	3	
			$p(1) > p(2)$	2	
		$p(0) > p(1)$	$p(1) \leqq p(2)$	1	
			$p(1) > p(2)$	0	

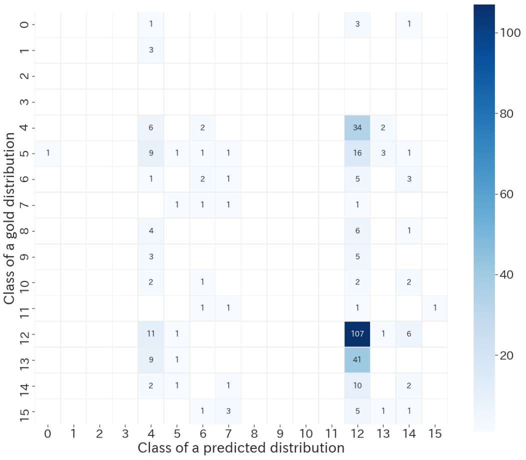


Figure 2 The heat map of the number of dialogues for A-score (20% of the training dataset)

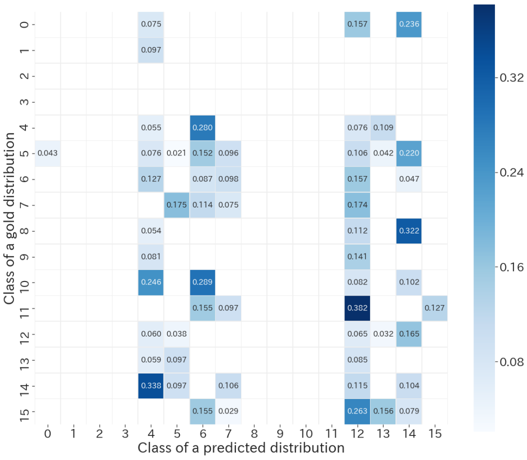


Figure 3 The heat map of Average NMD for A-score (20% of the training dataset)

a loss function considering ordinal bins. We apply the loss function showed in Eq. (2) as a simple one.

$$L_{\text{diff}} = \frac{1}{B-1}\sum_{i}^{B-1}\{(p(i+1)-p(i))-(p^*(i+1)-p^*(i))\}^2 \tag{2}$$

We use the combination of loss functions as follows,

$$L = \alpha L_{\text{diff}} + (1-\alpha)L_{\text{ce}} \tag{3}$$

where α denotes the paramater to adjust the weight of L_{diff}.

We use Eq. (3) and the same neural network as the baseline BLSTM model of STC-3 excluding the loss function. In other words, when we set α to 0.0, the baseline and proposed model are same. We set α to 0.5 as this achieved lower (i.e., more effective) dialogue quality scores than 0.25 and 0.75 for the validation data.

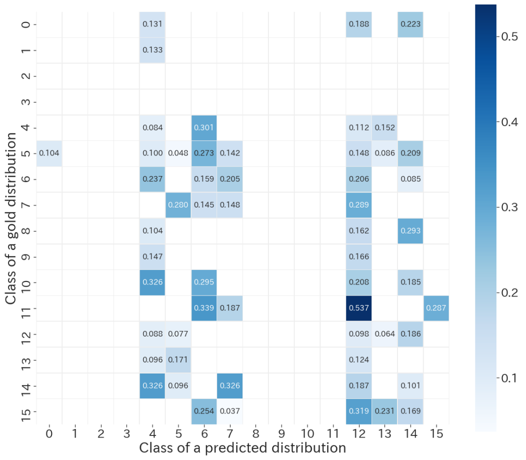


Figure 4 The heat map of Average RSNOD for A-score (20% of the training dataset)

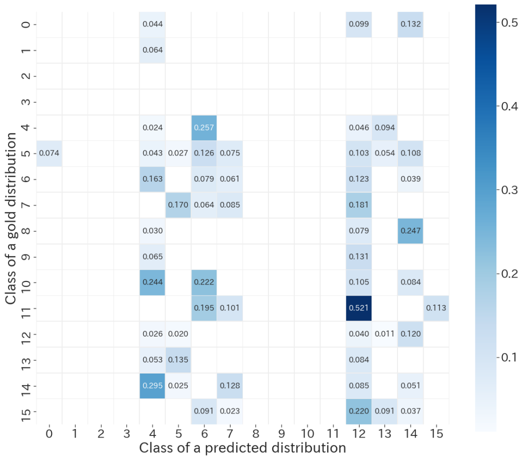


Figure 5 The heat map of Average JSD for A-score (20% of the training dataset)

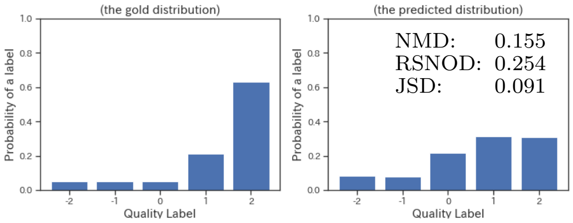


Figure 6 A Distribution Pair Example 1

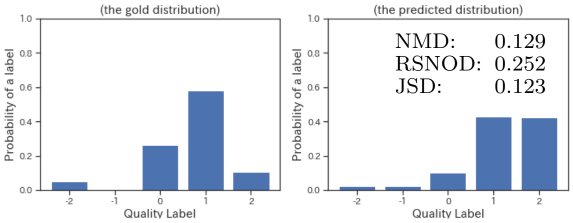


Figure 7 A Distribution Pair Example 2

5 Experiment

5.1 Dataset

We used the training dataset of STC-3 for the Chinese and English STC-3 DQ subtask. In the main experiment, we divide the dataset into the training, validation and testing data in the ratio of 7 : 2 : 1.

5.2 Training

As an example, we show the training loss and RSNOD as validation loss of each type of quality score and mean RSNOD over these types for English data in Figure 8 whose x-axis denotes the number of global steps of training. The training loss is scaled linearly to plot it close to the others. '•' denotes the minimum score of mean RSNOD over types of quality score for the validation data. We use the models which are trained until the global step of '•' to predict the distributions of the testing data.

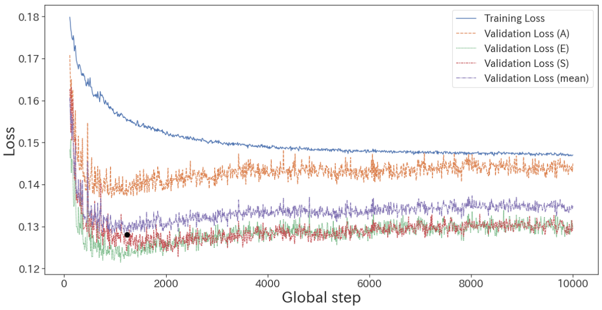


Figure 8 The training and validation loss

5.3 Result

As evaluation measures, we use NMD and RSNOD (the lower the better). We show the evaluation scores for STC-3 DQ subtask in terms of A-score, S-score and E-score of each language in Tables 2 to 7 where "Baseline" denotes the baseline BLSTM model of STC-3, "Diff+" denotes our proposed model and "only Diff" denotes our model which use α of 1.0. In Tables 2 to 7, the top scores are shown in bold. We conducted randomised Tukey HSD tests with $B = 10,000$ trials [7]. For English dataset, any models do not statistically significantly outperformed another at $\alpha = 0.05$, therefore we show p-values for only Chinese dataset in Tables 8 to 13. Effect sizes (i.e., standardized mean differences) based on one-way ANOVA (without replication) [9] are also shown. From the results with the evaluation measures for the DQ subtasks, it can be observed that:

- Our proposed model statistically significantly outperformed the baseline BLSTM model at $\alpha = 0.05$ in terms of RSNOD of E-score for Chinese dataset,
- Our proposed model outperformed the baseline BLSTM model in terms of every evaluation measure for Chinese dataset,
- Our proposed model outperformed the baseline BLSTM model in terms of NMD and RSNOD of S-score for English dataset.

Table 2 Chinese Results (A-score)

Model		Mean NMD	Mean RSNOD
Baseline	$(\alpha = 0.0)$	0.0776	0.1199
Diff+	$(\alpha = 0.5)$	**0.0762**	**0.1195**
only Diff	$(\alpha = 1.0)$	0.0812	0.1253

Table 3 Chinese Results (S-score)

Model		Mean NMD	Mean RSNOD
Baseline	$(\alpha = 0.0)$	0.0725	0.1174
Diff+	$(\alpha = 0.5)$	**0.0713**	**0.1146**
only Diff	$(\alpha = 1.0)$	0.0765	0.1221

Table 4 Chinese Results (E-score)

Model		Mean NMD	Mean RSNOD
Baseline	$(\alpha = 0.0)$	0.0756	0.1178
Diff+	$(\alpha = 0.5)$	**0.0719**	**0.1129**
only Diff	$(\alpha = 1.0)$	0.0798	0.1240

Table 5 English Results (A-score)

Model		Mean NMD	Mean RSNOD
Baseline	$(\alpha = 0.0)$	**0.0911**	**0.1307**
Diff+	$(\alpha = 0.5)$	0.0944	0.1333
only Diff	$(\alpha = 1.0)$	0.0939	0.1334

Table 6 English Results (S-score)

Model		Mean NMD	Mean RSNOD
Baseline	$(\alpha = 0.0)$	0.0853	0.1301
Diff+	$(\alpha = 0.5)$	**0.0844**	**0.1295**
only Diff	$(\alpha = 1.0)$	0.0863	0.1314

Table 7 English Results (E-score)

Model		Mean NMD	Mean RSNOD
Baseline	$(\alpha = 0.0)$	**0.0835**	**0.1256**
Diff+	$(\alpha = 0.5)$	0.0843	0.1265
only Diff	$(\alpha = 1.0)$	0.0856	0.1275

Table 8 Statistical significance in terms of NMD (Chinese, A-score)

	Diff+	only Diff
Baseline	$p = 0.747(0.0257)$	$p = 0.183(-0.0632)$
Diff+	-	$p = 0.023(-0.0889)$

Table 9 Statistical significance in terms of RSNOD (Chinese, A-score)

	Diff+	only Diff
Baseline	$p = 1.000(0.0050)$	$p = 0.014(-0.0791)$
Diff+	-	$p = 0.000(-0.0842)$

Table 10 Statistical significance in terms of NMD (Chinese, S-score)

	Diff+	only Diff
Baseline	$p = 0.718(0.0237)$	$p = 0.073(-0.0793)$
Diff+	-	$p = 0.000(-0.1031)$

Table 11 Statistical significance in terms of RSNOD (Chinese, S-score)

	Diff+	only Diff
Baseline	$p = 0.213(0.0407)$	$p = 0.016(-0.0670)$
Diff+	-	$p = 0.016(-0.1077)$

Table 12 Statistical significance in terms of NMD (Chinese, E-score)

	Diff+	only Diff
Baseline	$p = 0.085(0.0785)$	$p = 0.029(-0.0888)$
Diff+	-	$p = 0.000(-0.1673)$

Table 13 Statistical significance in terms of RSNOD (Chinese, E-score)

	Diff+	only Diff
Baseline	$p = 0.032(0.0811)$	$p = 0.006(-0.1043)$
Diff+	-	$p = 0.000(-0.1854)$

6 Discussion

Our proposed loss function improved the baseline model for Chinese dataset; however, it did not improve the baseline model for English dataset. The texts of English dataset were translated manually from Chinese dataset and the same gold labels were used to calculate gold distributions of dialogues. Therefore, we think the causes are derived from the architecture of the baseline model. We need to know the differences between the baseline model for Chinese and English dataset. For example, the pre-trained word embedding matrix for Chinese and English dataset differ. We have to analyze the vocabulary cover rate of each word embedding matrix for each dataset.

In Table 5, the order of models in terms of mean NMD and mean RSNOD differ; namely, "Diff+" was defeated by "only Diff". When we select a model in terms of global step of training, we use mean RSNOD over types of quality score described in section 5. 2. We have to analyze how to select a model using the validation data.

In this study, we simply add two loss functions in Eq. (3). We did not plot two training losses separately. Moreover, we can update parameters of a model using each loss function alternately. We have to analyze how to combine two loss functions.

7 Conclusion and Future Work

We tackled the STC-3 DQ subtask and in this subtask, the evaluation measures, i.e., NMD and RSNOD, which consider ordinal probability bins are used, therefore, we proposed the loss function which considers adjacent probability bins. We compared the baseline model of STC-3 DQ subtask to our proposed model which utilize our proposed loss function and our proposed model outperformed the baseline model in terms of every measure for Chinese dataset. Moreover, our proposed model statistically significantly outperformed the baseline model at $\alpha = 0.05$ in terms of RSNOD of E-score for Chinese dataset.

In the future, we analyze the differences between the architecture of the baseline model for Chinese and English dataset in order to make our proposed loss function work for English dataset. We also do experiments using the official test dataset of STC-3 DQ subtask.

References

[1] Higashinaka, R., Funakoshi, K., Inaba, M., Tsunomori, Y., Takahashi, T. and Kaji, N.: Overview of Dialogue Breakdown Detection Challenge 3, *Proceedings of Dialog System Technology Challenge 6 (DSTC6) Workshop* (2017).

[2] Hochreiter, S. and Schmidhuber, J.: Long Short-term Memory, *Neural computation*, Vol. 9, No. 8, pp. 1735–1780 (1997).

[3] Huang, M., Zeng, Z., Kato, S. and Sakai, T.: NTCIR-14 Short Text Conversation Task (STC-3), `http://sakailab.com/ntcir14stc3/` (2018).

[4] Janocha, K. and Czarnecki, W.: On Loss Functions for Deep Neural Networks in Classification (2017).

[5] Perez, J., Boureau, Y.-L. and Bordes, A.: Dialog System & Technology Challenge 6 Overview of Track 1 - End-to-End Goal-Oriented Dialog learning, *Proceedings of Dialog System Technology Challenge 6 (DSTC6) Workshop* (2017).

[6] Rubinstein, R.: The Cross-Entropy Method for Combinatorial and Continuous Optimization, *Methodology And Computing In Applied Probability*, Vol. 1, No. 2, pp. 127–190 (1999).

[7] Sakai, T.: Metrics, Statistics, Tests, *PROMISE Winter School 2013: Bridging between Information Retrieval and Databases (LNCS 8173)*, pp. 116–163 (2014).

[8] Sakai, T.: Comparing Two Binned Probability Distributions for Information Access Evaluation, *Proceedings of The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pp. 1073–1076 (2018).

[9] Sakai, T.: *Laboratory experiments in information retrieval: Sample sizes, effect sizes, and statistical power*, Springer (2018). `https://link.springer.com/book/10.1007/978-981-13-1199-4`.

[10] Schuster, M. and Paliwal, K. K.: Bidirectional Recurrent Neural Networks, *IEEE Trans. Signal Processing*, Vol. 45, No. 11, pp. 2673–2681 (1997).

[11] Zeng, Z., Luo, C., Shang, L., Li, H. and Sakai, T.: Test Collections and Measures for Evaluating Customer-Helpdesk Dialogues, *Proceedings of The 8th International Workshop on Evaluating Information Access*, pp. 1–9 (2017).